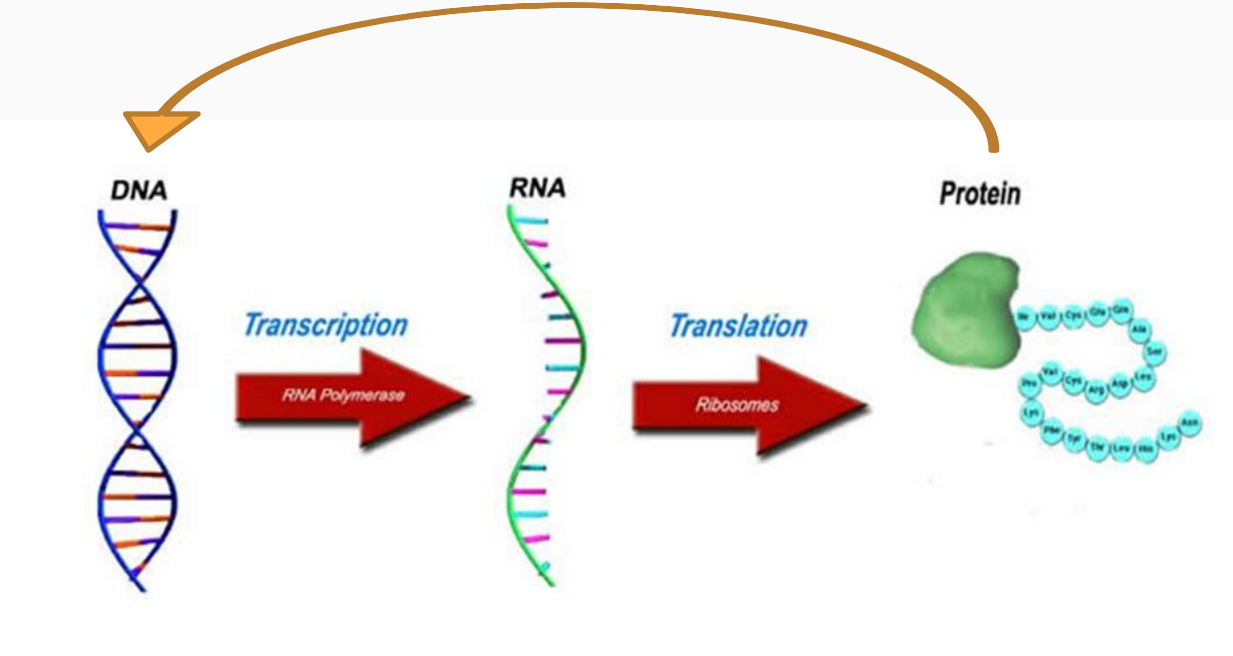
ChIP-seq analysis

Acknowledgements

Much of the content of this lecture is from:

- Furey (2012) – ChIP-seq and beyond
- Park (2009) – ChIP-seq – advantages + challenges
- Landt et al. (2012) – ChIP-seq guidelines + practices

Central Dogma of Biology

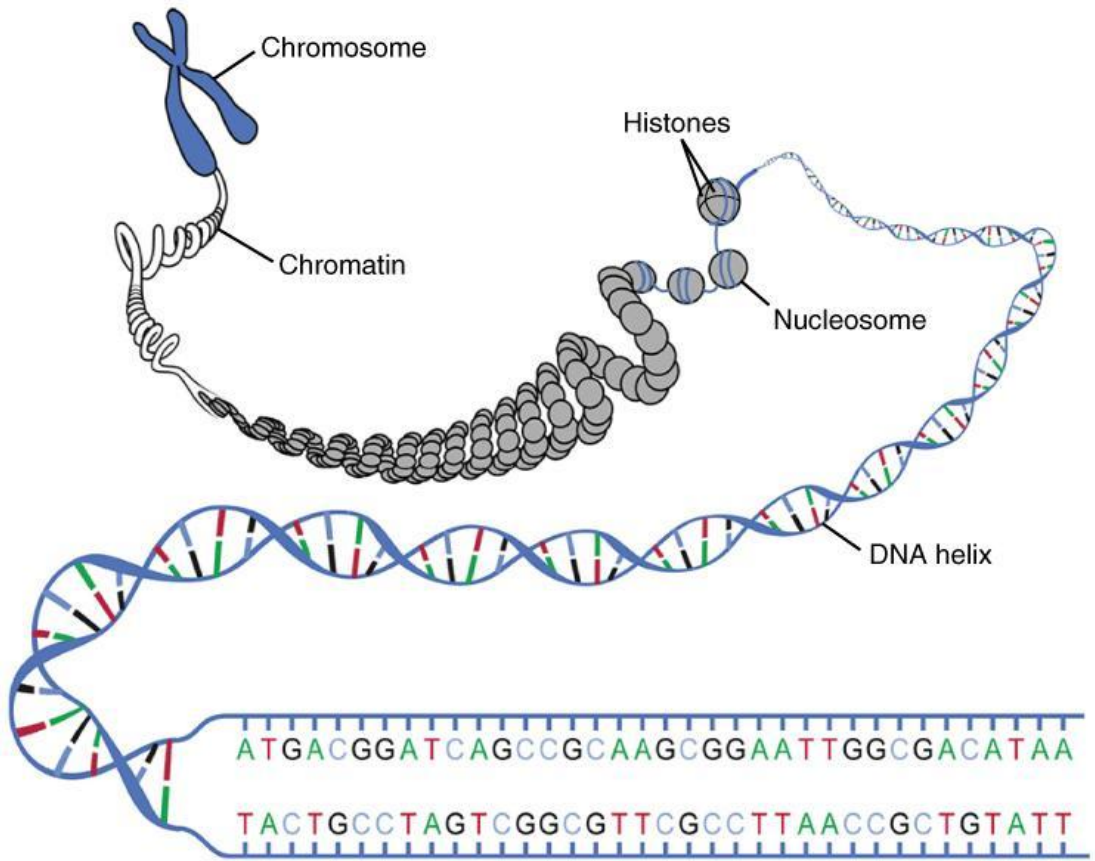


Some **proteins** can bind DNA to influence how genes are expressed

ChIP-seq

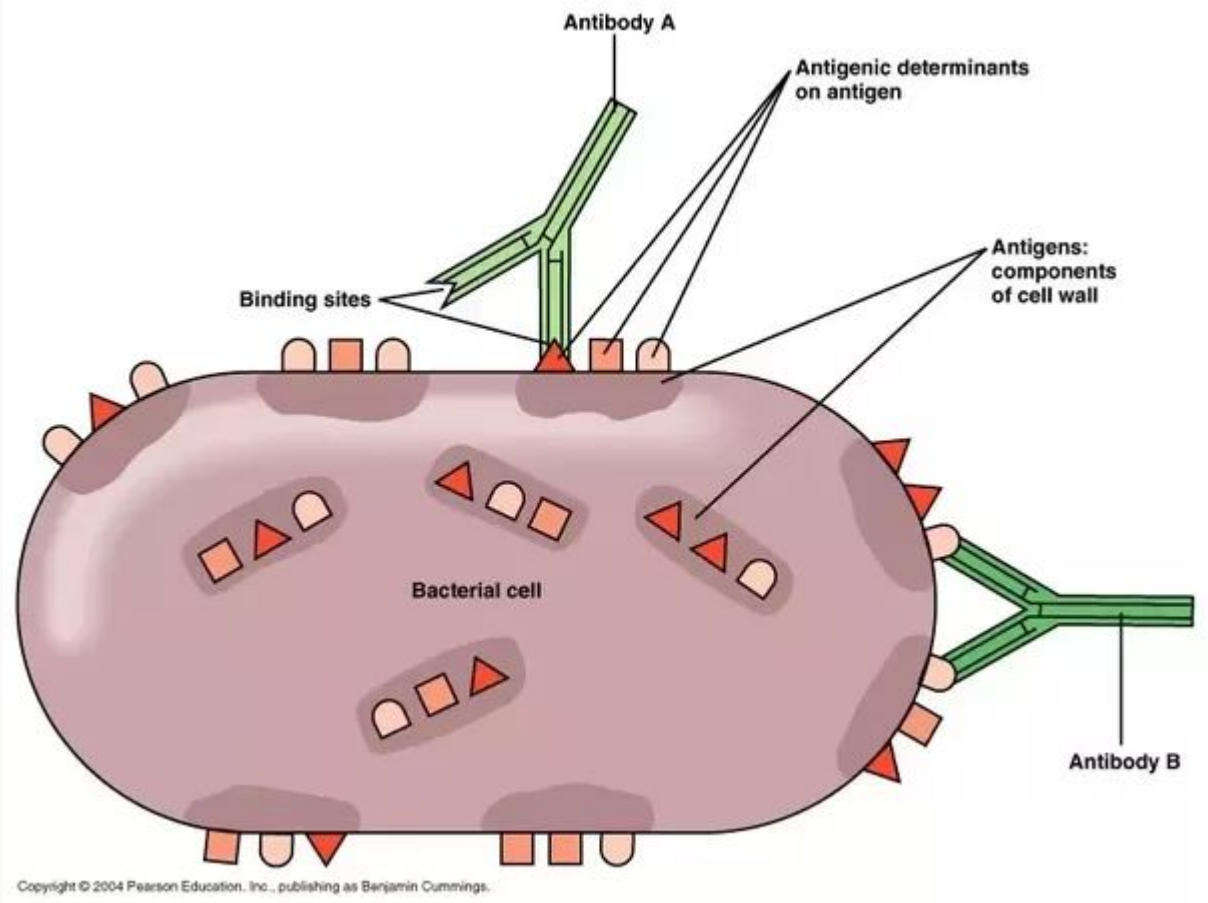
- **Chromatin immunoprecipitation** followed by high-throughput **sequencing**
- Assays the genome-wide locations of a **single protein** (bound to DNA) or a **single histone modification**

What is chromatin?



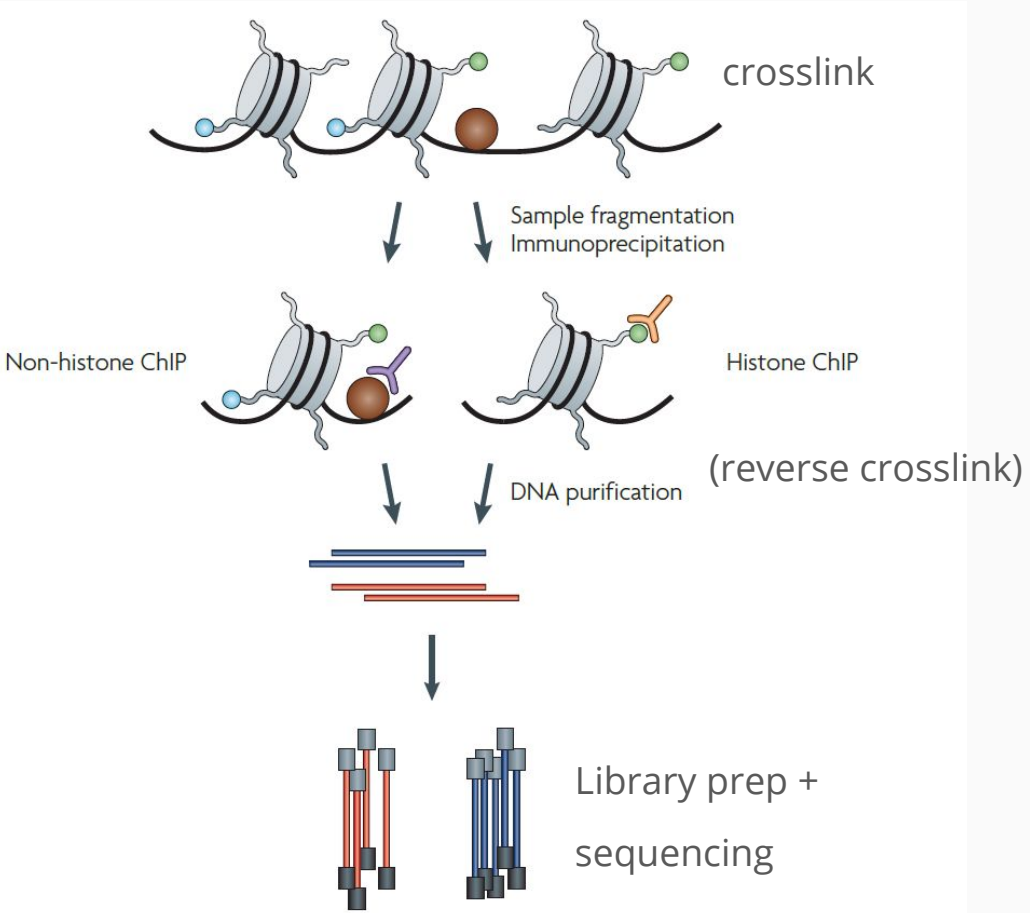
- Complex of macromolecules (DNA, protein, RNA)
- **Packages DNA** into compact shape
- Prevents DNA damage
- Controls gene expression, DNA replication

What is immunoprecipitation?



- **Antibodies** are used to immunoprecipitate proteins
- **Antibodies** bind in a (mostly) specific way to their **antigen**
- Used by the **immune system** to neutralize **pathogens**
- ChIP-seq uses antibodies raised against proteins

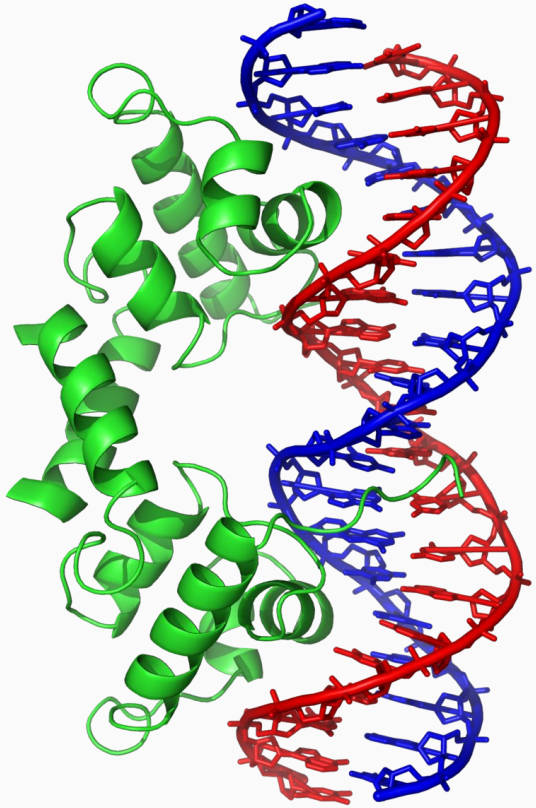
ChIP-seq protocol (very brief)



GOAL:

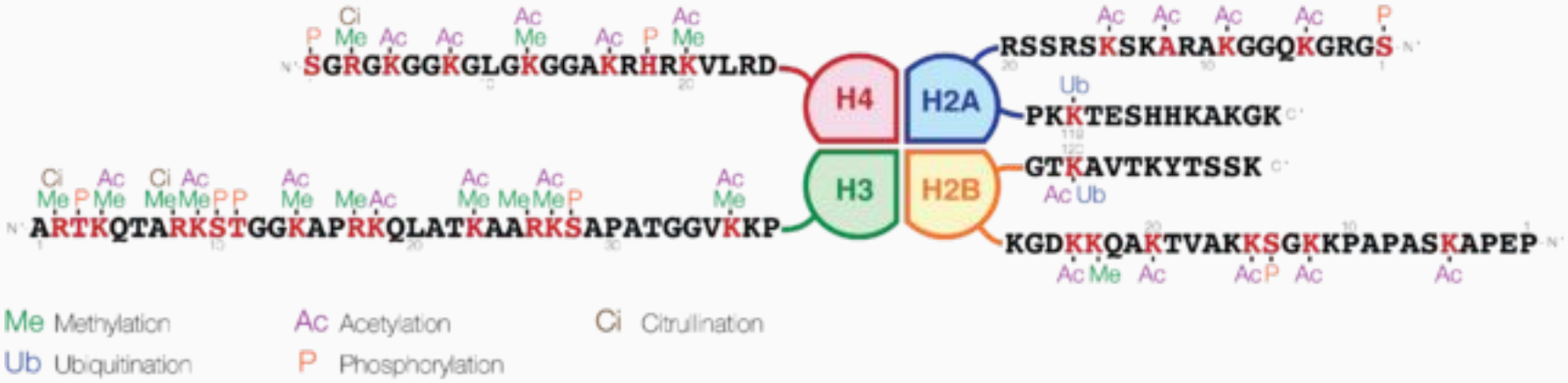
Determine **genomic DNA** associated with a given **protein** or **histone modification**

Why study protein binding to DNA?



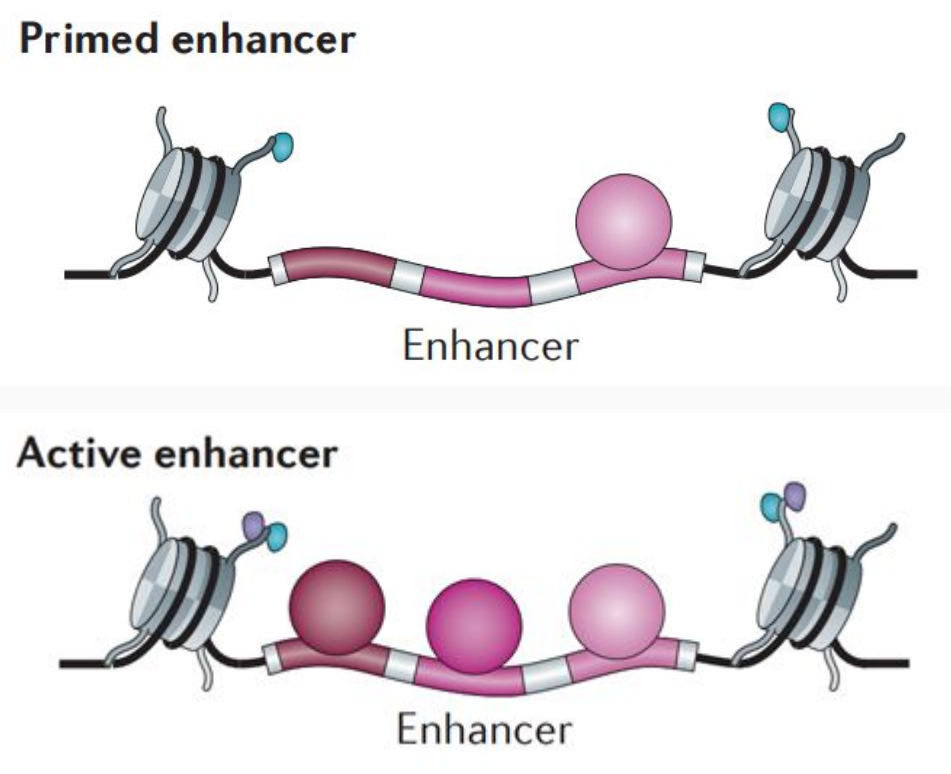
- **Transcription factors** (TFs) affect how genes are regulated
- DNA binding proteins (such as CTCF and cohesin) regulate the 3D structure of DNA

Why study histone modifications?

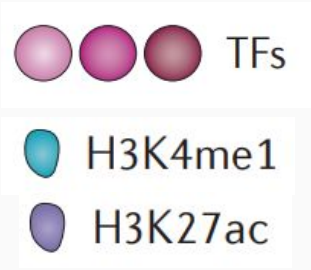


- **Combinations** of chemical modifications to the histone tails correlate with **regulatory activities**
- Referred to as the **“histone code”**

Integration of protein and histone ChIP-seq

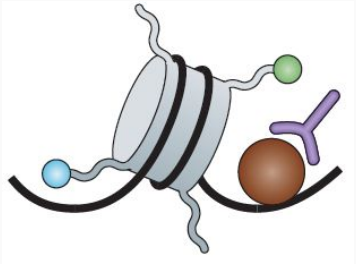


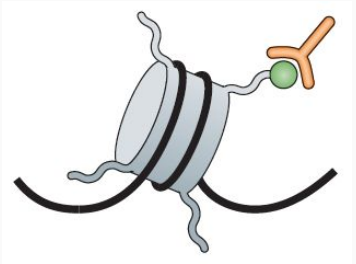
- ChIP-seq assays only 1 thing at a time
- Integration of several **proteins** and **histone modifications** provides more insight



Research Questions

Protein

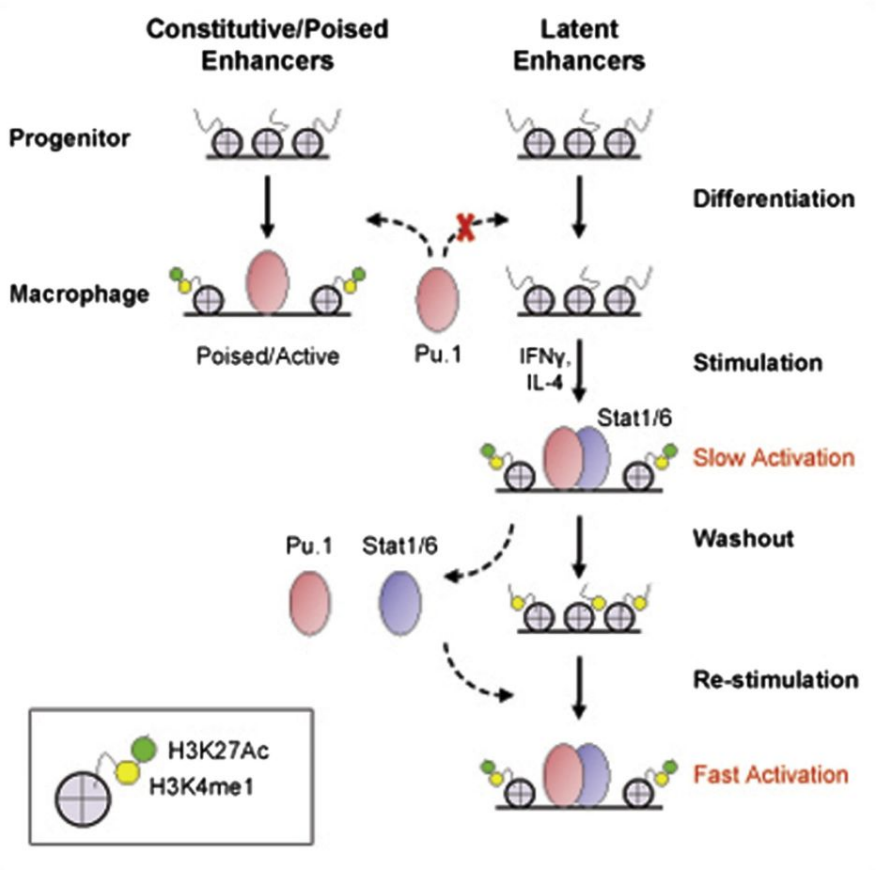




Histone mods

- DNA motif discovery – Which **DNA sequences** does my **protein** like to bind to OR which binding sequences correlate with a **histone modification**?
- Conserved/differential **protein binding** OR **histone modification** across conditions (time points, cell types, species, treatments)
- Genes (and gene sets) under regulation by a given **protein** or **histone modification**

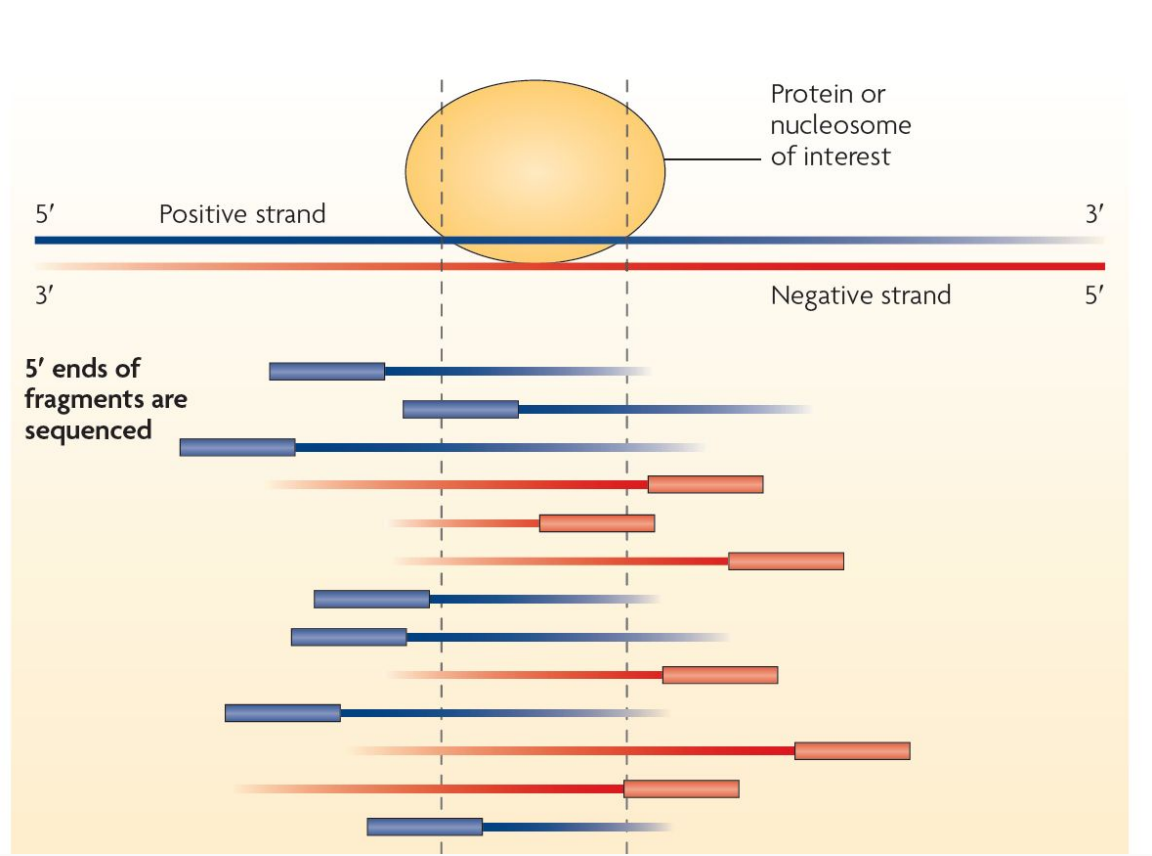
ChIP-seq study example



- Ostuni et al. (2013)
- Enhancer repertoire expanded during immune response
- Enhancers did not return to original state post-stimulus (epigenetic memory)
- Response upon restimulation was stronger and faster

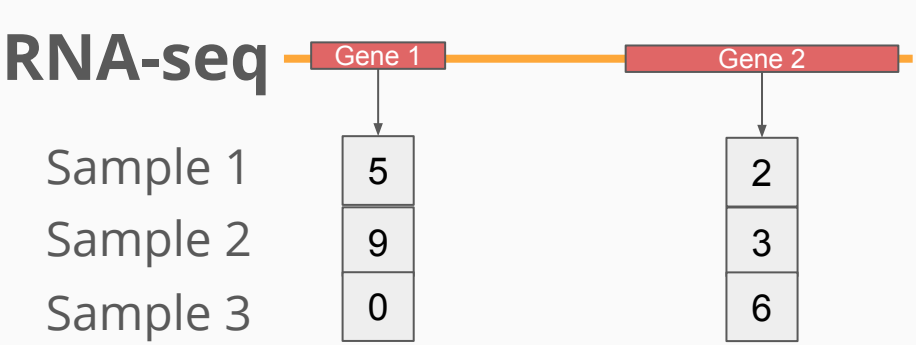
ChIP-seq Analysis Methods

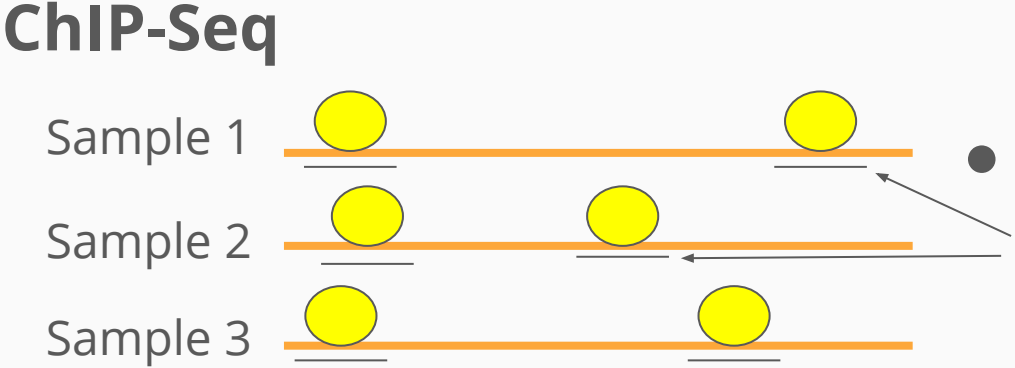
Preliminary Analysis Goal



- Define **where** your protein is binding or where histone modifications are occurring
- Inferred on a reference genome based on **short reads**

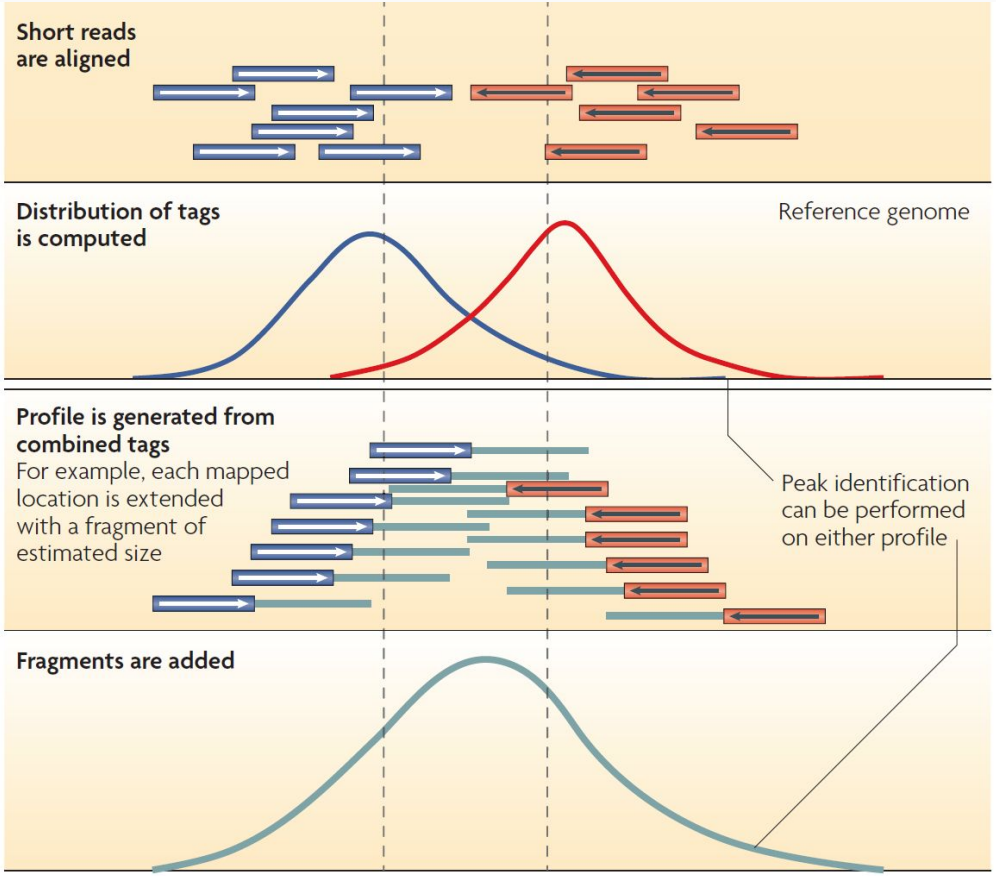
RNA-Seq vs ChIP-Seq: A key difference





- RNA-seq can use same gene annotation for each experiment
- Proteins can bind anywhere in the genome
- ChIP-seq features are experiment-specific
- Define features (called **peaks**) as part of the analysis pipeline

Defining ChIP-seq peaks

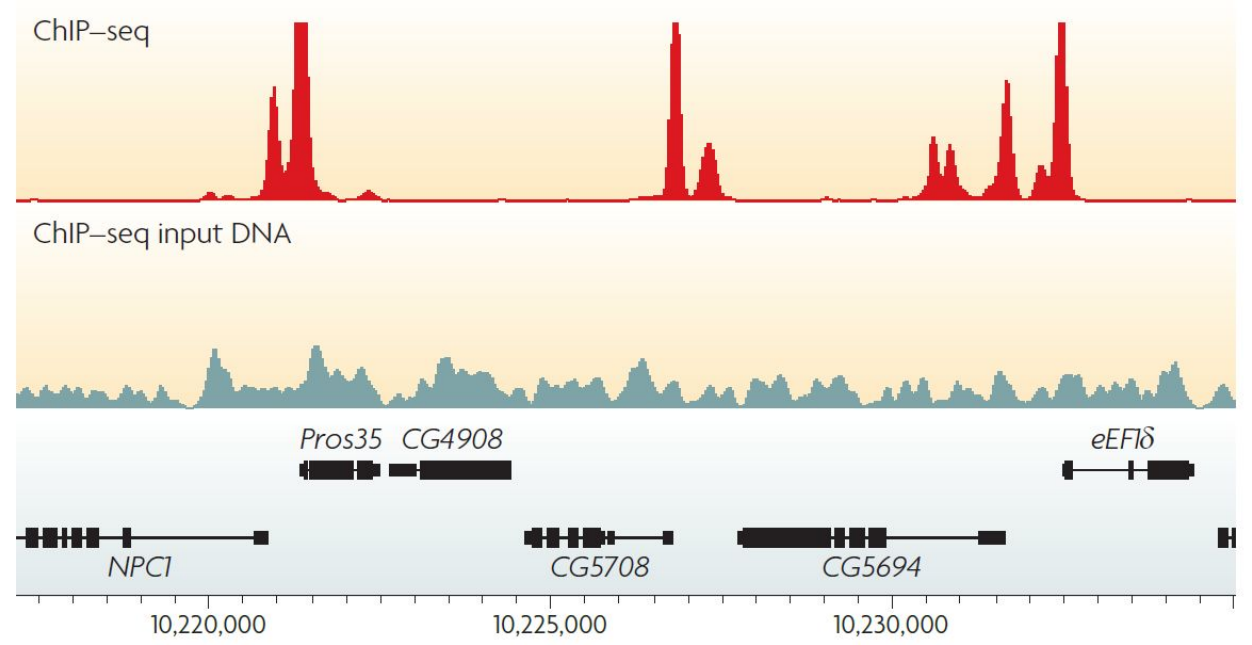


- **Peaks** are areas where read mapping is **enriched** compared to a **control** experiment
- Software exists to automate peak finding
- Popular programs include MACS2, GEM, HOMER, SPP

Controls for ChIP-seq

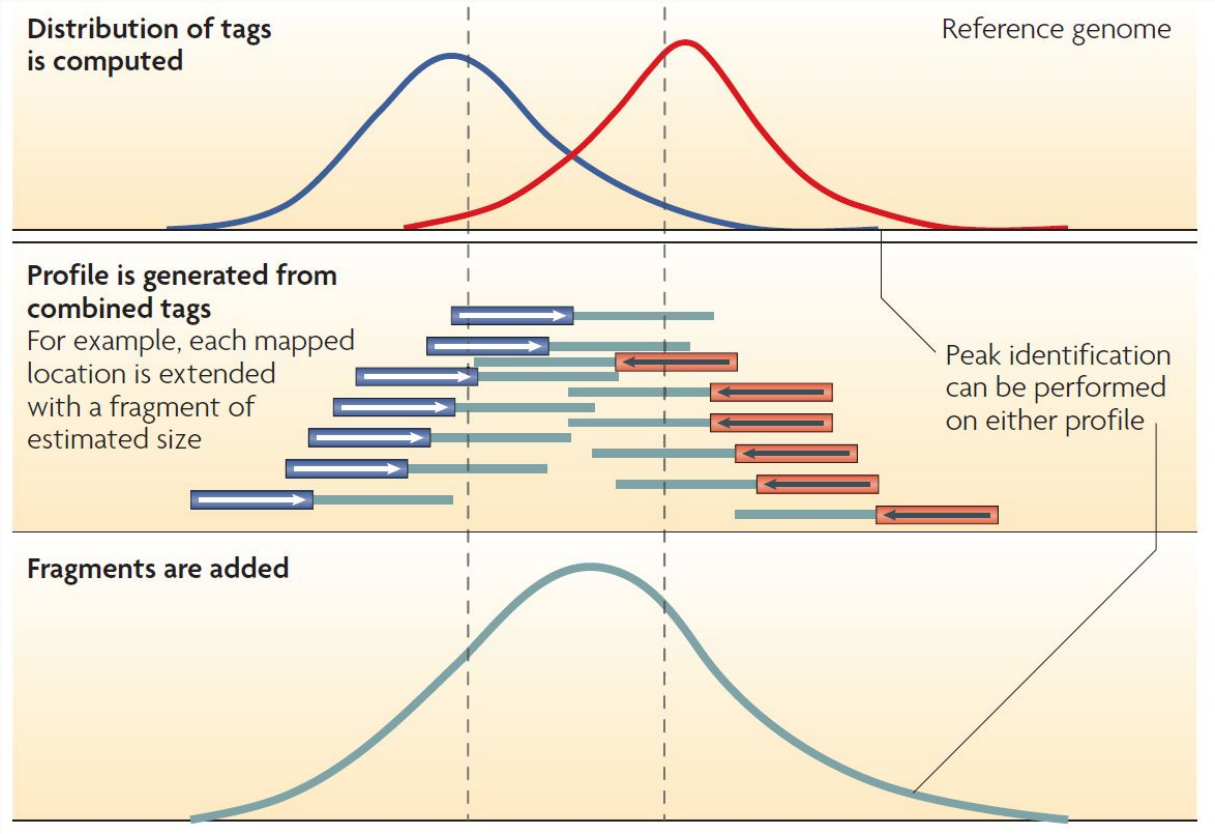
- **Input DNA** : A portion of DNA sample removed before immunoprecipitation
- **Mock IP** : DNA obtained from a fake IP performed without antibodies
- **IgG** : DNA from a non-specific IP using antibody against protein not involved in DNA binding
- Usually 1 is performed and most common is **input** which accounts for technical biases

ChIP-seq vs. input DNA



Input allows for correcting bias in variable solubility, shearing, and amplification during experiments

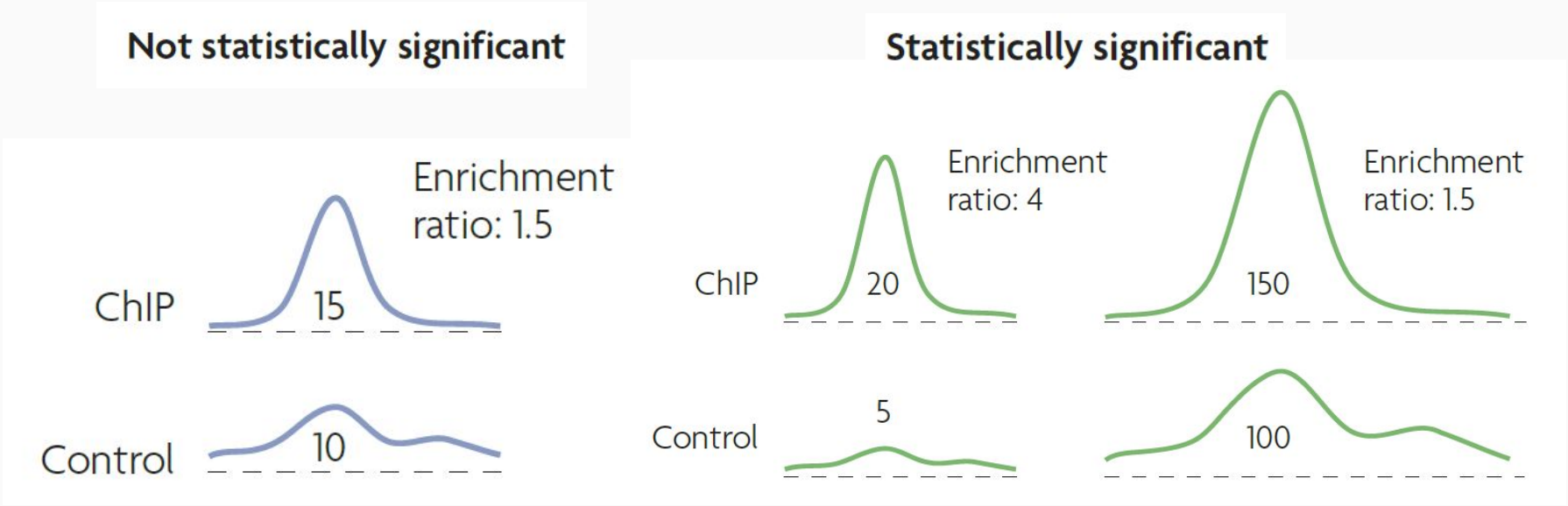
How does a peak caller work?



- Walks along the genome to identify enriched regions
- Estimates fragment size to extend reads into profile

Scoring peaks (general example)

- Poisson model for tag distribution accounts for **ratio** as well as **absolute tag number**

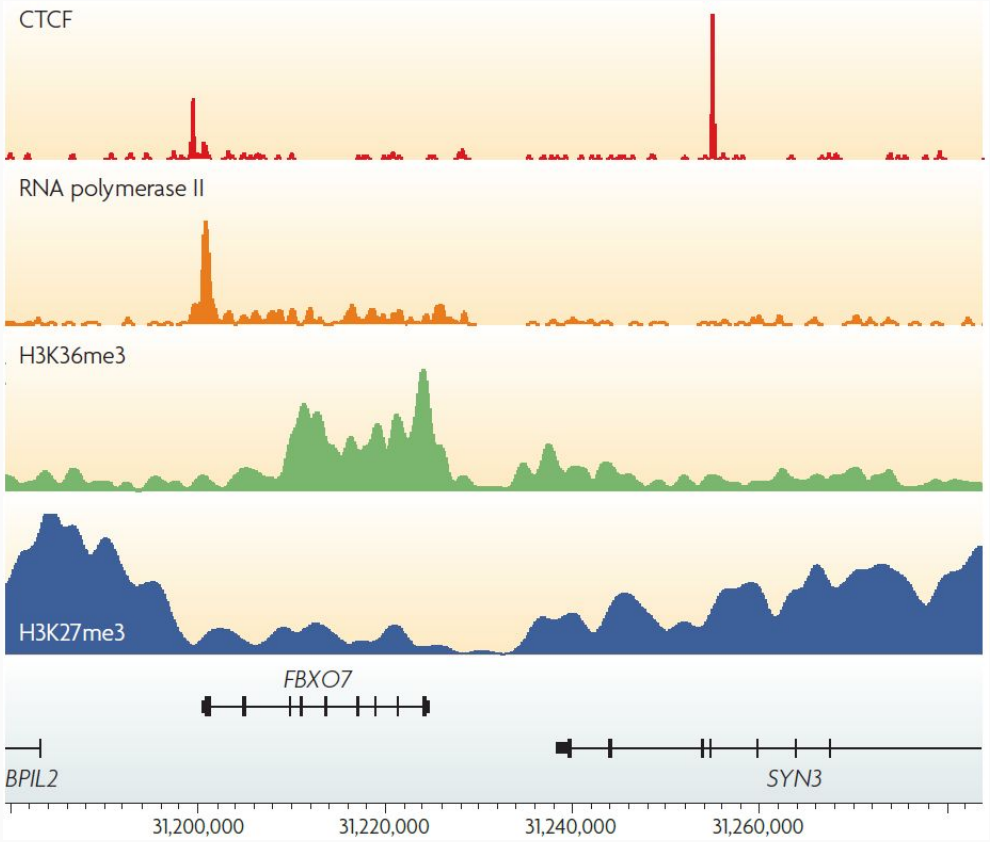


Significance of a Peak

- Statistical significance formally measured using **false discovery rate (FDR)**
- **FDR** is expected proportion of **incorrectly identified** sites among those found to be significant
- Can be measured by **swapping** input with ChIP sample and identifying false peaks
- The **q value** of a peak is the minimum FDR at which the peak is deemed significant
- Analogous to **p value** for a single hypothesis test

Peak calling for TF vs. histone mark

- Histone ChIP-seq produces much broader regions of enrichment
- Peak callers usually have a “histone” option or set of “broad” parameters if needed



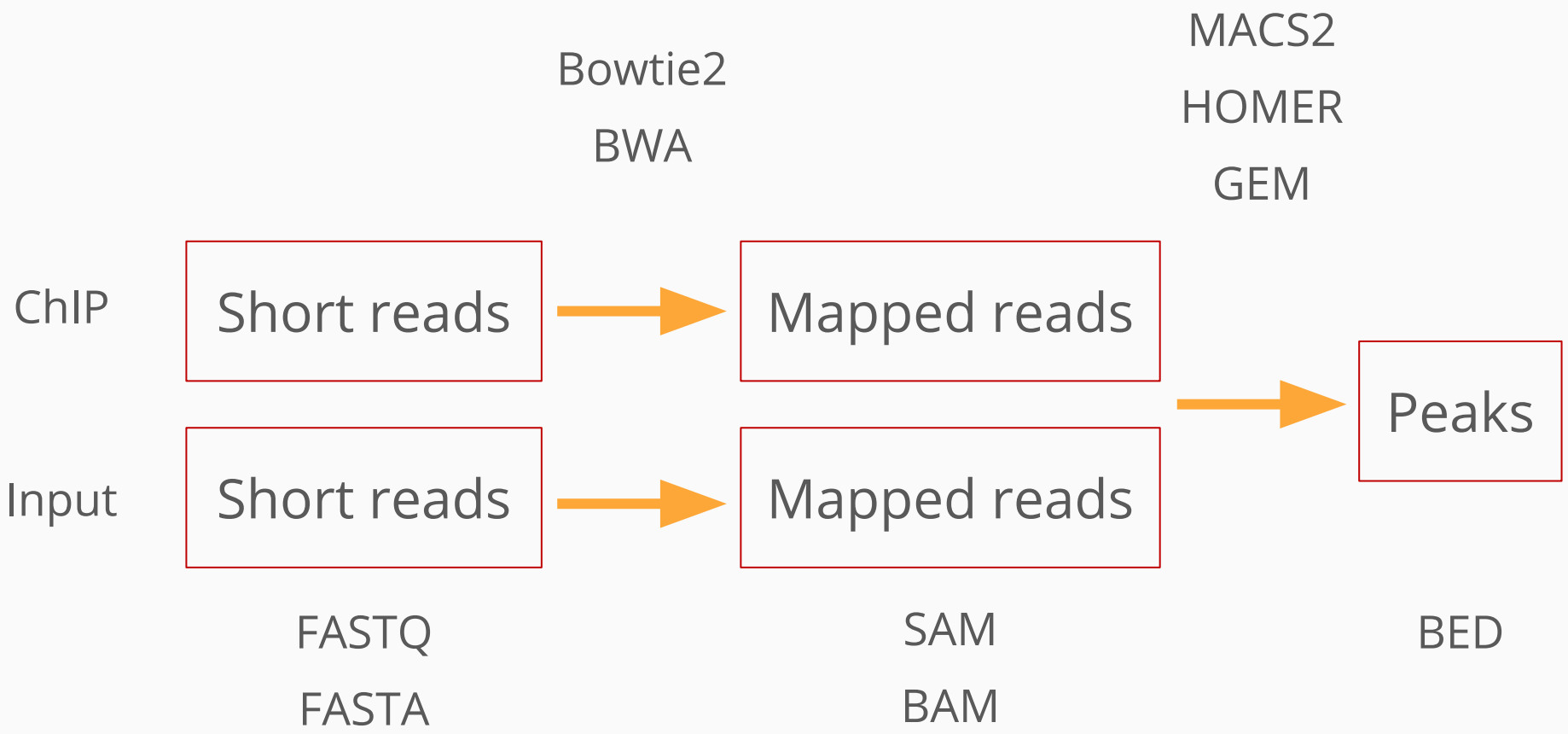
Output from peak calling

- Took a while to get here…
- List of **genomic loci** where either your **protein is bound** OR your **histone is modified**) – usually BED format

1	chr1	96412	96770	MACS2_DEF_CEBPA_93_BR2_peak_1	397	.	14.25609	42.47579	39.79834
2	chr1	241017	241233	MACS2_DEF_CEBPA_93_BR2_peak_2	90	.	6.06885	11.31148	9.04904 35
3	chr1	540931	541139	MACS2_DEF_CEBPA_93_BR2_peak_3	64	.	5.15916	8.59532	6.40643 35
4	chr1	715007	715301	MACS2_DEF_CEBPA_93_BR2_peak_4	346	.	13.04316	37.29144	34.66133
5	chr1	743138	743340	MACS2_DEF_CEBPA_93_BR2_peak_5	81	.	5.76562	10.38251	8.14303 55
6	chr1	748125	748451	MACS2_DEF_CEBPA_93_BR2_peak_6	359	.	13.34639	38.57281	35.93097
7	chr1	786996	787198	MACS2_DEF_CEBPA_93_BR2_peak_7	40	.	4.24947	6.11293	4.01359 33
8	chr1	893416	893666	MACS2_DEF_CEBPA_93_BR2_peak_8	107	.	6.09043	13.01581	10.71854 99
9	chr1	911556	911888	MACS2_DEF_CEBPA_93_BR2_peak_9	49	.	3.61846	7.09102	4.95216 176
10	chr1	926244	926498	MACS2_DEF_CEBPA_93_BR2_peak_10	97	.	4.76906	12.00287	9.72625 97
11	chr1	936083	936636	MACS2_DEF_CEBPA_93_BR2_peak_11	1072	.	16.02792	110.42943	107.275
12	chr1	944064	944475	MACS2_DEF_CEBPA_93_BR2_peak_12	586	.	11.83670	61.44181	58.60657
13	chr1	948668	948913	MACS2_DEF_CEBPA_93_BR2_peak_13	21	.	2.79661	4.16110	2.16002 65
14	chr1	963260	963755	MACS2_DEF_CEBPA_93_BR2_peak_14	249	.	10.61732	27.42960	24.90880

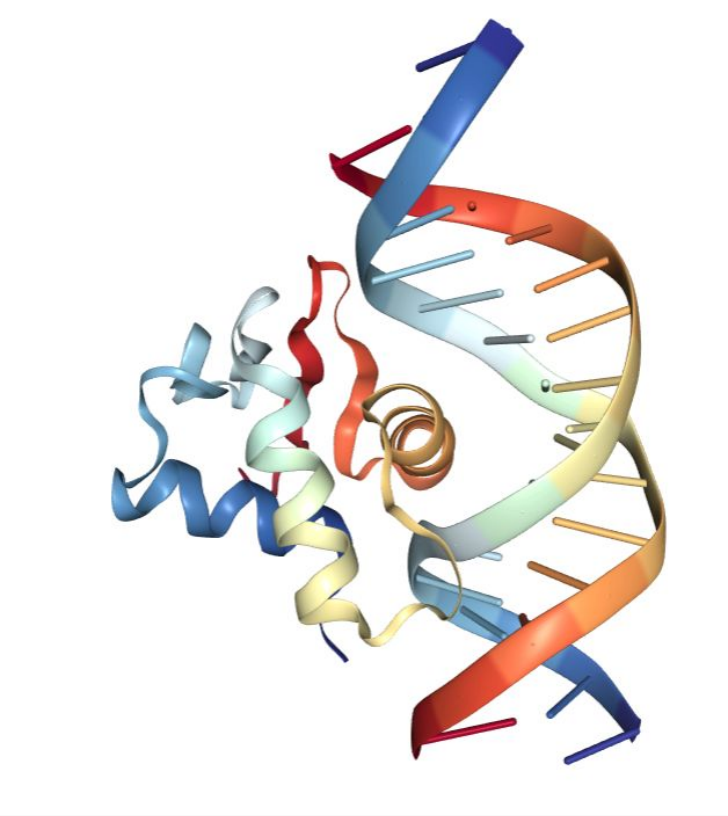
- Peak numbers vary wildly by protein, organism, etc.

Typical analysis workflow



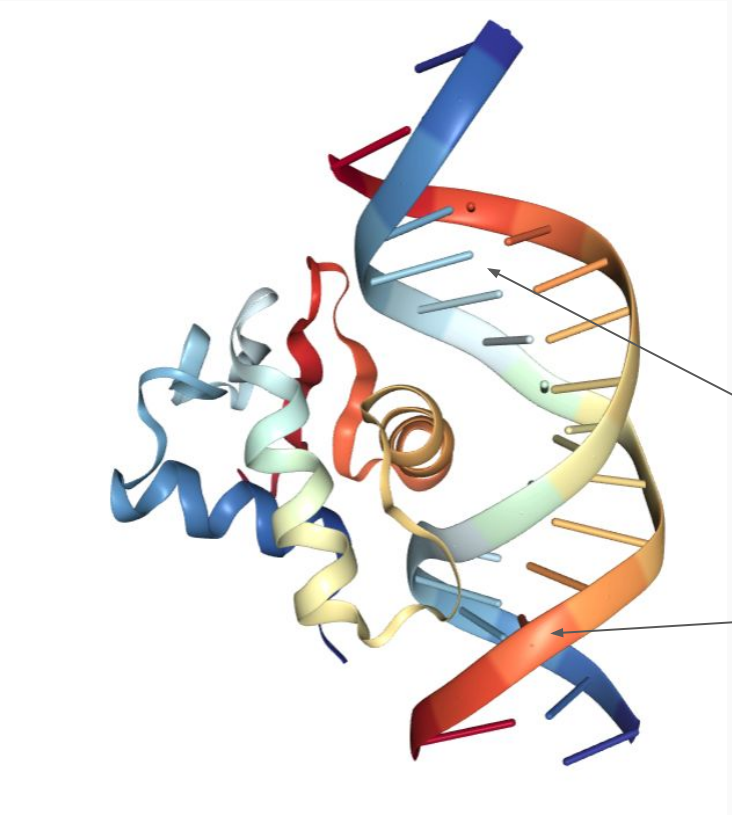
Functional Characterization

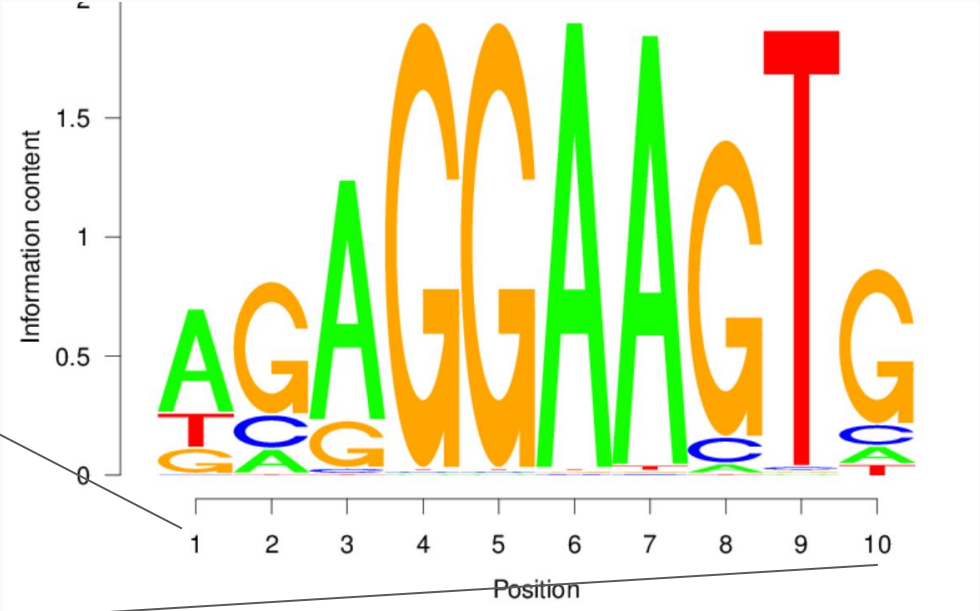
Where and how is my protein binding?



- **Peaks** are areas where read mapping is **enriched** compared to a **control** experiment (~300bp)
- Actual **binding sites** (for proteins) are 8-12bp
- Binding site can be inferred using **motifs and motif analysis**

What is a DNA-binding motif?

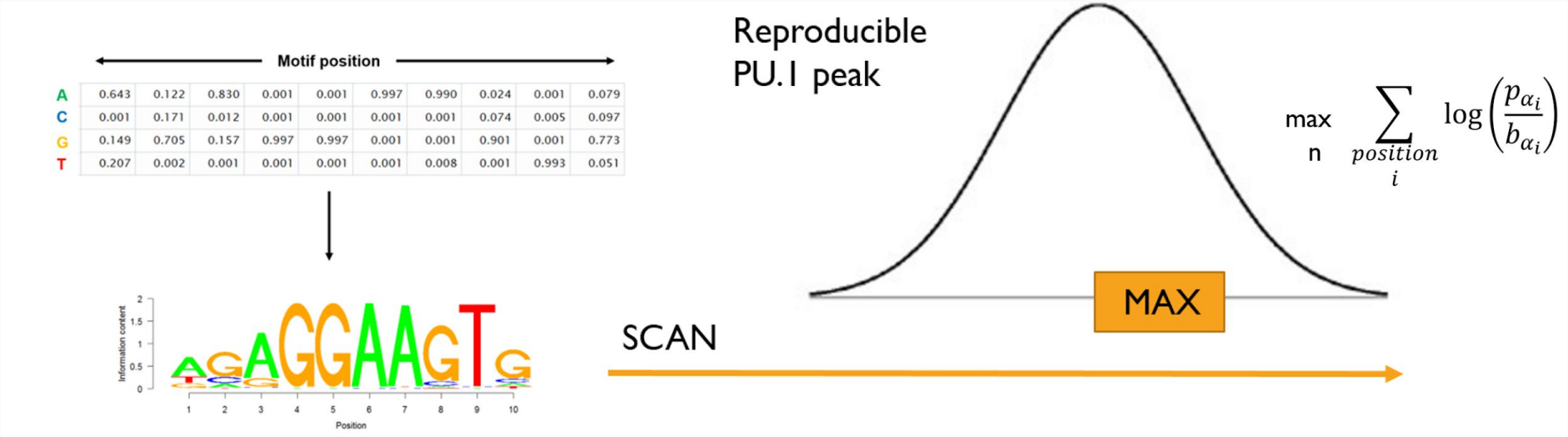




Motif position

A	0.643	0.122	0.830	0.001	0.001	0.997	0.990	0.024	0.001	0.079
C	0.001	0.171	0.012	0.001	0.001	0.001	0.001	0.074	0.005	0.097
G	0.149	0.705	0.157	0.997	0.997	0.001	0.001	0.901	0.001	0.773
T	0.207	0.002	0.001	0.001	0.001	0.001	0.008	0.001	0.993	0.051

Motif scanning (scoring)



- “Scan” for binding sites using probability model
- Ask at each position in peak “how likely is it that this is a binding site and not some random sequence?”
- Motif occurrences typically are located near peak summits

Motif scoring example

- How likely is it that this is a binding site and not some random sequence?

	T	G	G	G	G	A	A	G	T	G
A	0.643	0.122	0.830	0.001	0.001	0.997	0.990	0.024	0.001	0.079
C	0.001	0.171	0.012	0.001	0.001	0.001	0.001	0.074	0.005	0.097
G	0.149	0.705	0.157	0.997	0.997	0.001	0.001	0.901	0.001	0.773
T	0.207	0.002	0.001	0.001	0.001	0.001	0.008	0.001	0.993	0.051

- Pr (binding site) = 0.207 x 0.705 x 0.830 ...
- Pr (random seq) = 0.250 x 0.250 x 0.250 ...

What if I don’t know the binding site?

2 general approaches:

- **Motif enrichment analysis**: Scan a **library of known motifs** against your peaks (and a background) to determine which motifs are most enriched
- **De novo motif finding**: learns **new motifs** using expectation/maximization (MEME) or k-mer based approaches (HOMER)
- If chipping a protein previously done, **both motif analyses should yield similar results**

Example Homer report (enrichment)

Homer Known Motif Enrichment Results

Homer *de novo* Motif Results
Gene Ontology Enrichment Results
Known Motif Enrichment Results (txt file)
Total Target Sequences = 15213, Total Background Sequences = 34081

Rank	Motif	Name	P-value	log P-pvalue	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Motif File	PDF
1		NFkB-p65(RHD)/GM12787-p65-ChIP-Seq/Homer	1e-1707	-3.931e+03	3855.0	25.34%	1506.4	4.42%	motif file (matrix)	pdf
2		CEBP(bZIP)/CEBPb-ChIP-Seq/Homer	1e-1310	-3.018e+03	3423.0	22.50%	1551.0	4.55%	motif file (matrix)	pdf
3		PU.1(ETS)/ThioMac-PU.1-ChIP-Seq/Homer	1e-1288	-2.967e+03	3413.0	22.44%	1569.7	4.61%	motif file (matrix)	pdf
4		AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq/Homer	1e-947	-2.183e+03	3251.0	21.37%	1900.8	5.58%	motif file (matrix)	pdf
5		NFkB-p65-Rel(RHD)/LPS-exp/Homer	1e-936	-2.157e+03	1163.0	7.65%	166.2	0.49%	motif file (matrix)	pdf
6		ETS1(ETS)/Jurkat-ETS1-ChIP-Seq/Homer	1e-885	-2.039e+03	4447.0	29.23%	3568.0	10.47%	motif file (matrix)	pdf

Example Homer report (de novo)

Homer *de novo* Motif Results

Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP
* - possible false positive

Rank	Motif	P-value	log P-pvalue	Best Match/Details	Motif File
1		1.407e-35	-8.025e+01	NFkB-p65/GM12787-p65-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
2		1.371e-14	-3.192e+01	ISRE/ThioMac-LPS-exp/HOMER More Information | Similar Motifs Found	motif file (matrix)
3		3.312e-14	-3.104e+01	CRE/Promoter/Homer More Information | Similar Motifs Found	motif file (matrix)

What is my protein doing?

- Integration with **RNA-seq** data – You can do pathway/ontology EA using nearest gene **(careful!)**
- **Differential binding** – Very similar to RNA-seq (even uses the same software – genomic loci instead of genes)
    - DESeq2, DiffBind, etc.
- Integration with **other ChIP-seq experiments** – Does my protein bind enhancers? Repressed regions? Co-bind with other proteins?
    - Will talk more about these in **integrative genomics**